A Probabilistic Model for Multi-Contestant Races

By Konstantinos Gakis[*]
Panos Pardalos[†]
Chang-Hwan Choi[‡]
Jae-Hyeon Park[+]

Predictions of sports games have been recognized as an important area of study for its economic significance. The majority of models for such games cover two-player games and the resulting championships or study individual players or teams and their resulting comparative position. However, many sports involve race-type multi-contestant games, which are more complex in modeling. In this paper we outline the difficulties associated with the study of such races

Keywords: *Ranking, Joint Exponential Distribution, Order Statistics, Multi-contestant Races*

Introduction

Predictions of sports games have been recognized as an important area of study for its economic significance. The majority of models for such games cover two-player games and the resulting championships or study individual players or teams and their resulting comparative position. Most popular sports involve predictions of the final score, by fans and bookmakers, making necessary the creation of a performance index for athletes or teams.

Most of the research and the literature have been concerned with two contestant games, in many cases with a score, where one wins and one loses or there can be a tie. The models developed try to predict whether the result will be a win, loss, or a tie, and other even going as far as guessing a score-line for the matches. A whole variety of parameters can be considered to make predictions for a match, such as the previous results for both teams (scores, historical classification), previous meetings between the teams, and so on, even a simulation engine can be developed, to use factors related to training, players in team and other physiological and psychological agents.

[*] Adjunct Assistant Professor, Dept. of Industrial & Systems Engineering, University of Florida, USA.

[†] Distinguished Professor, Dept. of Industrial & Systems Engineering, University of Florida, USA.

[‡] Post-Doctoral Researcher, Kinesmetrics Laboratory & Center for Performance and Sports, Analysis, Korea National Sport University, Korea.

[+] Associate Professor, Kinesmetrics Laboratory & Center for Performance and Sports Analysis, Korea National Sport University, Korea.

https://doi.org/10.30958/ajspo.3-2-1 doi=10.30958/ajspo.3-2-1

In several sports, such as football, tennis and basketball, ranking indices have been developed to better shape the classification of individual player qualities. It is common these days to rank sports performance of the athletes.

Ranking system is typically used in team sports such soccer, baseball, and basketball as well as in individual sports such as badminton, tennis, and tae-kwon-do. It derives from the curiosity of "Who is the best player?" We would like to be aware of a certain player's generalized performance ability even if players do not compete against each other. Different ranking system are used depending on sports teams or individual athletes. Examples include accumulative point system of International Federation of Association Football (FIFA) to rank national performance, Terry-Bradley model to rank players competing pairwise, Elo Systems of International Chess Federation (ICF) to calculate the players' relative skill. There are two basic tournament systems: round robin system and cup system. As tournaments generate the results of wins, losses, or ties mostly in individual tournaments, ranking models have been developed on the basis of individual games. About these sports, several methods were developed to predict match results. The most recent research is based on several machine learning techniques [2,6,9], but also statistical inference [1,3] and Page Rank method [8].

These methods use a relatively simple model that relies on a limited amount of data, for example by monitoring previous matches. Other kinds of sports are not based on one-to-one matches and the results is a classification. These are characterized as "multi-contestant races" or "games." For such games, prediction models become more complex and usually they are based on physiological information about athletes [7]. However, often it is necessary to predict results based only on easily available data, such as previous results, keeping a reliable model.

The main aim of this paper is to build a theoretical stochastic model for predicting results of multi-contestant games. For this reason we consider the case when the times of the athletes is exponentially distributed. Although this is not realistic in most of the known cases, the assumption leads to the built-up of a model that provides valuable insight to the problem.

The initial assumption is that the athletes' performance is independent of each other and their performance capacity is steady over time. We are proposing an approach that can allow us to relax the assumption of independence and introduce some correlation between the performances of athletes.

General Description

The outcome of a multi-contestant race or game is basically a ranking problem. Let's assume that we have seven athletes competing in a race. The outcome will depend on the individual performance of each athlete. Let's say that the race outcomes are *running times*, in which case the winner is the athlete that comes first, i.e. the one that achieves the minimum of the times. So, let $X_i, i = 1, \dots, m$ (assuming *m* athletes are competing) be the times (which are random variables) achieved by the athletes in a particular game. Then, the probability of a particular outcome would be expressed as, for e.g. *m* = 4

$$\Pr\{X_1 < X_2 < X_3 < X_4\}$$

or the probability of athlete 1 being the first

$$\Pr\{\text{athlete 1 terminates first}\} = \Pr\{X_1 < X_2, X_3, X_4\}$$

The simplest case would be to assume that running times are independently distributed. Then the fact that several athletes compete in a race is really irrelevant. What would matter is just individual performance. From experience we know that this is not true. The fact that athletes are competing against each other makes a difference in the running times and thus the assumption of independence cannot be realistic. We clearly need a more complex assumption. Since for the prediction we only need probability for the relative performance, i.e. the ranking and not the absolute times, we can assume as an approximation that the simultaneous presence of many athletes in the same competition just reduces the time scale but not the final ranking probabilities. So, when our sample from races is large and we have results for all possible combinations of athletes from the pool, we can build a prediction model that would yield good results in terms of predictive power. If this is the case and we have races of *m* athletes out of a pool of *n* > *m*, then to express the probability of a particular outcome it would suffice to know the relative strength of each athlete in the form of a *dominance* vector $D = \{d_1, \dots, d_n\}$ that not only *ranks* the athletes, but quantifies the dominance as well.

We can deduce the relative strength indirectly from their relative strength compared to other athletes. In this case we make the assumption that:

Let “:>” denote “better than”

When $A{:>}B \wedge B{:>}C \Rightarrow A{:>}C$ (1)

If this were so, then a simple ranking would provide a sound basis for prediction.

Our analysis of bicycling races data, however, suggests that the above proposition cannot be a realistic assumption. Thus, we have to seek for a way to introduce the element of dependence in outcomes depending on the composition of the racing athletes group. When studying real sets of data from multi-contestant races we encounter three significant difficulties that complicate the problem of developing a model and estimating its parameter values:

- The first difficulty is the fact that in many case we have the rank and not the actual race time performance available.

- The second difficulty is related to the fact that we do not have historical data about contests between all pairs of athletes, thus we cannot even answer the question “who is better, athlete A or athlete B?”
- The third difficulty stems from the fact that there are several cases that paired comparisons seemingly lead to inconsistent results, such as when $A{:>}B \wedge B{:>}C$ but $C{:>}A$.

In our model we propose some ways of overcoming these difficulties.

Independently Distributed Exponential Race Times

The Basic Problem

We have a pool of $N, N = 1,2, \dots$ athletes competing in groups of $M, M = 1,2, \dots, N$ athletes each time. We have data results from $K, K \gg N$, races.

In this first approach we assume independently and identically distributed negative exponential random variables (“expo r.v.’s”) for the individual race time of each athlete. Say the individual parameters of the distributions are $i, i = 1, \dots, N$.

With knowledge of the actual ranks, without knowing the actual race times, we can obtain an estimate of the relative rank. We know that when comparing two iid expo r.v.’s, then the probability of 1 coming ahead of 2 is

$$p_{12} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

With a large data set we can have the actual proportion of times athlete 1 came ahead of athlete 2, $\hat{p}_{12}$ and thus set

$$\lambda_2 = \frac{1 - \hat{p}_{12}}{\hat{p}_{12}} \lambda_1$$

An immediate difficulty comes from the fact that if we add a third player we will have

$$\lambda_3 = \frac{1 - \hat{p}_{13}}{\hat{p}_{13}} \lambda_1$$

and

$$\lambda_3 = \frac{1 - \hat{p}_{23}}{\hat{p}_{23}} \lambda_2$$

or

$$\lambda_3 = \frac{1 - \hat{p}_{23}}{\hat{p}_{23}} \frac{1 - \hat{p}_{12}}{\hat{p}_{12}} \lambda_1$$

That would entail that

$$\frac{1-\hat{p}_{13}}{\hat{p}_{13}} = \frac{1-\hat{p}_{23}}{\hat{p}_{23}}\frac{1-\hat{p}_{12}}{\hat{p}_{12}}$$

which in general will not be the case!

Thus, we have to perform a maximum likelihood estimation (MLE) for the parameters based on the distribution function of the order statistics of the exponential distribution (for a discussion of the problem, the reader is referred to [11]).

Note that to have a solution to the problem we need a sample that does not have $\hat{p}_{ij}, j = 0 \text{ or } 1, i \neq j$ for any pair of *i,j*. This approach allows us to rank the athletes and thus make predictions based on the estimated parameters expressed only relative to one of the parameters. It would be simpler if we picked as basis the best or fastest athlete and express all others' ranks as fractions of that best athlete.

Problems Related to This Approach

Although the approach we presented can yield practical results in all cases where $\hat{p}_{ij} = 0 \text{ or } 1, \text{for } i \neq j,$ there is a series of theoretical problems associated:

- The data set for every pair may not be of the same in size, thus introducing a serious element of imbalance in the significance of the estimators. The effect of sample size should be investigated in order to be aware about the possible shortcomings of the model.
- As the mix of athletes is different in every case, what exactly is the maximum likelihood function? Is the solution of ignoring non-participating athletes a good solution?
- Although the problem of estimation of the maximum likelihood estimators is simple for a small number of observations and athletes, it may become very tedious for larger numbers. The study of these problems is part of our current research.

Non-Independently Distributed Exponential Race Times

A next step is to relax the assumption of independence. This relaxation seems to better fit actual race data as the relative performance of athletes is differs to a significant degree depending on the mix of athletes in each race.

Gumbel [10] presented a series of jointly distributed non-independent random variables with negative exponential marginal distributions. One such joint distribution of two random variables X_1 and X_2 is defined by the (bivariate) probability density function:

$$f(x,y) = e^{-(x+y)}[1+\alpha(2e^{-x}-1)(2e^{-y}-1)]$$

where

$$-1 \leq \alpha \leq 1$$

Gumbel [9] shows that for this distribution the correlation coefficient is $\rho = \alpha/4$.

We generalize this distribution for $m = 1,2, \dots$:

$$f(x_1, x_2, \dots, x_m) = e^{-\sum_{i=1}^{m} x_i}[1 + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \alpha_{ij}(2e^{-x_i} - 1)(2e^{-x_j} - 1)] \tag{2}$$

where again we require that

$$-1 \leq \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \alpha_{ij} \leq 1$$

We call $\alpha_{ij} = \alpha_{ji}$ the interaction coefficient factor among athlete *i* and athlete *j*. This distribution offers a useful tool to study races with exactly the same participants. In order to study the broader problem of races with *m* athletes from a pool of *n* athletes, we can introduce a set of new parameters, the dominance factor $i, i = 1,2, \dots, n$, where *n* is the total number of athletes in the pool of athletes. We also define the interaction factor α_{ij}, which expresses the relative interaction among two athletes in general and not in the context of a particular race. For particular races, the factor is scaled to a coefficient taking into account the make-up of the race. Also, let for every race define

$$\alpha_{ij} = \frac{\delta_{ij}}{\delta^*}$$

$$\delta^* = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \delta_{ij}$$

$$\lambda^* = \prod_{i=1}^{m} \lambda_i$$

Then we can redefine Expression (2) as

$$f(x_1, x_2, \dots, x_m) = \lambda^* e^{-\sum_{i=1}^{m} \lambda_i x_i}[1 + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{\delta_{ij}}{\delta^*} (2e^{-\lambda_i x_i} - 1)(2e^{-\lambda_j x_j} - 1)] \tag{3}$$

Our current research focuses on the study of the order statistics of this distribution, which will enable the application of parameter estimation techniques such as Maximum Likelihood Estimation to determine the relative values of the parameters $\lambda_i, i = 1,2, \dots, n$ and $\delta_{ij}, i = 1,2, \dots, n$ and $j = 1,2, \dots, n$.

Further Steps

In this paper we have stated briefly the problem of multi-contestant races and we presented a formulation of the problem that does not appear in the literature so far.

We explained why the problem is very complex and it requires large sets of data. Existing methods do not allow the estimation of the parameters and new methods, which will necessarily rely on heavy iterative computing algorithms, will have to be developed.

Next steps will include estimation of the parameters for the data-set and further theoretical elaboration of the probability law of the joint distribution of the random variables that were presented.

Acknowledgment

This work was partially supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A2A2028551).

References

[1] Esters, Irvin G., and F. Uttenbach Richard. "Utility of team indices for predicting end of season ranking in two national polls." *Journal of Sport Behavior* 18 (1995): 216-224.

[2] Baio, Gianluca, and Marta Blangiardo. "Bayesian hierarchical model for the prediction of football results." *Journal of Applied Statistics* 37.2 (2010): 253-264.

[3] Boulier, Bryan L., and Herman O. Stekler. "Predicting the outcomes of National Football League games." *International Journal of Forecasting* 19.2 (2003): 257-270.

[4] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine." *Computer networks and ISDN systems* 30.1 (1998): 107-117.

[5] Deb, Kalyanmoy. *Multi-objective optimization using evolutionary algorithms.* Vol. 16. John Wiley & Sons, 2001.

[6] Joseph, A., Norman E. Fenton, and Martin Neil. "Predicting football results using Bayesian nets and other machine learning techniques." *Knowledge-Based Systems* 19.7 (2006): 544-553.

[7] Lamberts, Robert P. "Predicting cycling performance in trained to elite male and female cyclists." *Int. J. Sports Physiol. Perform* (2013).

[8] Lazova, Verica, and Lasko Basnarkov. "PageRank Approach to Ranking National Football Teams." *arXiv preprint* arXiv:1503.01331 (2015).

[8] Min, Byungho, et al. "A compound framework for sports results prediction: A football case study." *Knowledge-Based Systems* 21.7 (2008): 551-562.

[9] Gumbel, E. J. “Bivariate exponential distributions.” *J. Amer. Statist. Assoc.* 55 (1960): 698-707.

[10] Ahsanullah, M., V. B. Nevzorov, M. Shakil, *An Introduction to Order Statistics,* Atlantis Press, 2013.